Building Causal Graphs from Medical Literature and Electronic Medical Records

Galia Nordon
Technion
Israel Institute of Technology
Haifa, Israel

Gideon Koren
Maccabi-Kahn
Institute of Research and Innovation
Tel-Aviv, Israel

Varda Shalev
Maccabi-Kahn
Institute of Research and Innovation
Tel-Aviv, Israel

Benny Kimelfeld
Technion
Israel Institute of Technology
Haifa, Israel

Uri Shalit
Technion
Israel Institute of Technology
Haifa, Israel

Kira Radinsky
Technion
Israel Institute of Technology
Haifa, Israel

Abstract

Large repositories of medical data, such as Electronic Medical Record (EMR) data, are recognized as promising sources for knowledge discovery. Effective analysis of such repositories often necessitate a thorough understanding of dependencies in the data. For example, if the patient age is ignored, then one might wrongly conclude a causal relationship between cataract and hypertension. Such confounding variables are often identified by causal graphs, where variables are connected by causal relationships. Current approaches to automatically building such graphs are based on text analysis over medical literature; yet, the result is typically a large graph of low precision. There are statistical methods for constructing causal graphs from observational data, but they are less suitable for dealing with a large number of covariates, which is the case in EMR data. Consequently, confounding variables are often identified by medical domain experts via a manual, expensive, and time-consuming process.

We present a novel approach for automatically constructing causal graphs between medical conditions. The first part is a novel graph-based method to better capture causal relationships implied by medical literature, especially in the presence of multiple causal factors. Yet even after using these advanced text-analysis methods, the text data still contains many weak or uncertain causal connections. Therefore, we construct a second graph for these terms based on an EMR repository of over 1.5M patients. We combine the two graphs, leaving only edges that have both medical-text-based and observational evidence. We examine several strategies to carry out our approach, and compare the precision of the resulting graphs using medical experts. Our results show a significant improvement in the precision of any of our methods compared to the state of the art.

1 Introduction

Electronic Medical Record (EMR) data is a powerful resource for discovering medical and other health-related knowledge. EMRs are being widely adopted for use in observational causal studies, where the causal effect of an intervention is sought (Avillach et al. 2012; Stuart et al. 2013; Casucci et al. 2017; Gottlieb et al. 2017). Crucial to observational studies is the question of which covariates may influence the studied effect. These covariates, called *confounders and mediators* (depending on whether they cause or are caused by the intervention), need to be taken into consideration when analyzing the data.

While ostensibly the large amount of covariates in EMR data is an advantage, in practice it often creates difficulties in observational studies. Standard methods for causal inference in observational studies, such as *propensity score adjustment or matching*, often fail when the number of covariates is large, e.g. in the thousands. This is due to the fact that it is very difficult to define good metrics in high-dimensions, and that propensity scores often show lack of overlap in high-dimensions (D'Amour et al. 2017). A common practice (Triantafillou et al. 2017) is therefore to look at a small set of confounders and mediators that are believed to be most relevant to the problem at hand. This set is usually built manually by experts. However, expert knowledge is sometimes limited and not fully reproducible, and moreover, it does not account for the full potential of EMRs to discover new knowledge. In this work, we focus on the challenge of automatically constructing useful causal graphs.

Several notions of health graphs have been proposed as means of capturing expert medical knowledge from the literature (Rotmensch et al. 2017; Goodwin and M. Harabagiu 2013). Yet, adapting these graphs for causal inference carries numerous challenges. It is difficult to identify confounders and mediators from disconnected pieces of causality knowledge extracted from text.

In this work, we propose a new approach to constructing a causal graph from observational data originating from EMR, while incorporating knowledge about interventional distributions gained through the countless experiments present in the medical literature (a repository of 27 million medical abstracts and citations from PubMed). Our approach consists of three main steps, which we describe next.

The first two steps represent a novel graph-based method to better capture causal relationships implied by medical literature while attempting to preserve causal context. In the first step we analyze PubMed sentences for causal relationships. We leverage the SemRep (Rindflesch and Fiszman 2003) extractor to retain edges that contain causal semantic predicates, e.g., "Mediastinal emphysema due to acute bronchial asthma." In the second step we perform *graph embeddings* of the derived terms, optimizing an objective that preserves local neighborhoods of nodes (Perozzi, Al-Rfou, and Skiena 2014; Tang et al. 2015). We do this in order to connect disconnected causal facts. We study recent graph embeddings techniques (Grover and Leskovec 2016) that extend the objective to optimize for embeddings that capture similarity in network neighborhoods. We show that inferring causal relations with semantic similarity over such graph embeddings improves significantly the proposed causal-graphs.

Yet, even after using the above text-analysis techniques, the text data still contains many weak or uncertain causal connections. Therefore, in the third step we use the relevant concepts extracted from the medical literature to construct another graph, this time using the EMR data. While the text graph is based on causal connections implied by natural language in the text, the EMR graph is undirected and contains only correlations. We use a *lack of correlation* in the EMR data as a criterion for pruning the dense text-derived graph. We note that our use of the observational data is different from approaches which attempt to address the much more difficult problem of fully constructing a causal graph from data, where the challenges are mainly because the data itself is often incomplete and the large number of variables that need to be accounted for (Hauser and Bühlmann 2015; Triantafillou and Tsamardinos 2015; Triantafillou et al. 2017).

We perform an extended experimental evaluation comparing several causal graph construction methods, that show the merit of each one of the steps as evaluated by medical experts. Our empirical results show significant precision gains of the resulting causal graph.

To the best of our knowledge, there is currently no publicly available database of causal medical condition relationships. We consider our work as the starting point of a collaborative effort of creating such a database. Our graph not only supplies the causal relations, but also the specific textual references that generated the edge (i.e., the *provenance* of the relationship), making the graph highly interpretable. We are making our results available for other researchers to use and contribute [1]. We especially encourage the addition of correlations obtained from EMRs reflecting diverse populations.

2 Related Work

Observational data has previously been used for constructing causal relations and graphs (Claassen and Heskes 2012; Triantafillou and Tsamardinos 2015; Triantafillou et al. 2017; Sachs et al. 2005). These methods measure independence and conditional independence between variables in the data and construct a causal graph accordingly. The main shortcoming of these methods is that they are limited in the number of variables that can be modeled. Rotmentch et al. (Rotmensch et al. 2017) used several probabilistic models to construct disease-symptom graphs based on EMRs, without considering medical literature. Observational data, such as EMR data, is often limited in its representation as it contains a short summary of symptoms and doctor's decisions, missing the theoretical medical knowledge behind them. Moreover, a causal relation such as disease A is caused by condition B is often considered irrelevant to the EMR record and the records will contain mention of both diseases without an annotation of the fact one is a result of the other. Hence, EMRs will produce useful correlations between conditions but the information of the nature of the correlation – is it causal or not – will be missing. Goodwin et al. (Goodwin and M. Harabagiu 2013) attempt to capture such hidden relations from physicians notes. They focus on assessing the level of physician belief in a medical condition rather than extracting causal relations. Finlayson et al. (Finlayson, LePendu, and Shah 2014) built a "graph of medicine" based on co-occurrence of medical concepts in clinical notes.

Medical literature, on the other hand, focuses on the theory and explanation of biomedical processes. It naturally contains explanations of processes and causation. Natural language processing (NLP) has been successfully used for extraction of relations in various domains (Radinsky and Davidovich 2012) and has also been applied to medical literature. Bui et al. (Bui et al. 2010) extracted causal relations of HIV drug resistance. SemRep (Rindflesch and Fiszman 2003) presented a comprehensive data set of predications describing subject-object relations from medical paper abstracts, amongst them are causation predicates.

3 Data Model and Repositories

Throughout the paper, we assume a set $\mathbf{D}$ of *medical conditions*. We extract relationships over $\mathbf{D}$ from two types of data repositories: textual data from PubMed and EMR data from a Maccabi Healthcare, Israel's second largest healthcare provider.

PubMed repository. PubMed is a search engine accessing all MEDLINE (Kilicoglu et al. 2012) citations and several other resources. It is a literary repository of over 27 million citations and abstracts of biomedical academic literature. As such, it represents detailed professional peer-reviewed medical knowledge. For the sake of simplicity, we refer to this repository simply as "PubMed" in the remainder of the paper.

We use SemRep (Rindflesch and Fiszman 2003) to extract semantic propositions from the MEDLINE text, as a basis for a text-based causal graph. At the initial stages of tokenization and part-of-speech identification, domain-specific noun phrases are identified and mapped to concepts in a specialized metathesaurus, based on the Unified Medical Language System, UMLS (Bodenreider 2004). In UMLS, each concept belongs to a predefined UMLS semantic type, and a semantic network of relations between concepts and types is defined. A rule-based algorithm maps semantic propositions between concepts to a set of predefined predicates such as "treats," "causes" and "is-a." The mapping is done using both syntactic rules and semantic constraints based on the UMLS semantic net. The resulting predicates are subject-predicate-object tuples, each extracted from a single sentence in a PubMed abstract. For a full description of the extraction process we refer the reader to (Rindflesch and Fiszman 2003).

In formal terms, we use SemRep to extract triples of the

[1] https://github.com/TechnionTDK/causal-graphs

form (d_1, p, d_2) where $d_1, d_2 \in \mathbf{D}$ are medical conditions, and p is a predicate relationship between medical conditions. Hence, we extract a *directed labeled graph* over medical conditions, represented as a bag of triples. The elements in $\mathbf{D}$ are identified as those having the UMLS type *Disease or Syndrome*.

EMR repository. We use Electronic Medical Records (EMRs) from Israel's second largest health care provider, serving more than two million patients, covering the years 2005-2010. The repository holds a complete medical history for each patient, including disease diagnosis codes used by clinicians. These codes are specific to the healthcare provider and are mapped to icd-9 codes. More specifically, the repository contains $115,000,000$ diagnoses out of $27,519$ possible diagnosis codes. Abstractly, we view this repository as a collection of pairs (q, d), stating that person q has been diagnosed with the condition $d \in \mathbf{D}$.

4 Building a Medical Condition Causal Graph

A medical condition causal graph is a directed graph where the nodes belong to $\mathbf{D}$, and edges represent causal relations between nodes—an edge $e = (d_1, d_2)$ where $d_1, d_2 \in \mathbf{D}$, states that d_1 causes d_2. In this section, we describe in detail our construction of the causal graph.

Our general approach is to build a graph from the PubMed data, and then to prune it using an EMR-based graph. Figure 1 illustrates the main steps of our algorithm. We first construct a text-based causal-graph from the textual repository (boxes (1)-(3)). Next, we reduce the graph using *medical condition embeddings* (boxes (4)-(5)), where medical conditions that share the same network community and/or similar roles have close vectors in a latent space of low dimension, using one of several embedding methods outlined further in this section. The inferred graph embeddings allow us to filter for each medical condition a set of potential causal neighbors (box (6)). We complete the construction by detecting the causal neighbors by leveraging correlations identified in the EMR repository (boxes (7)-(9)).

Building a Text-Based causal-graph

In Section 3, we described how we extract a directed labeled graph over medical conditions. This graph consists of triples (d_1, p, d_2) where d_1 and d_2 are medical conditions and p is a relationship between them. However, as a causal graph, these triples contain noise and misclassification due to errors in semantic misunderstandings. For example, the question: "Is Celiac disease caused by allergy to Gliadin?" is translated into a causal relation between Celiac and Gliadin. It may also contain errors due to inconsistent text. To overcome these drawbacks, we construct a more accurate directed graph according to these predicates, experimenting with several methods for text-edge construction, outlined below. In each method, we identify a collection of *causal* predicates p from UMLS: *causes, prevents, disrupts, inhibits, predisposes, produces*. A triple (d_1, p, d_2) is *causal* if p is a causal predicate. We present the following methods for the causal text-edge construction:

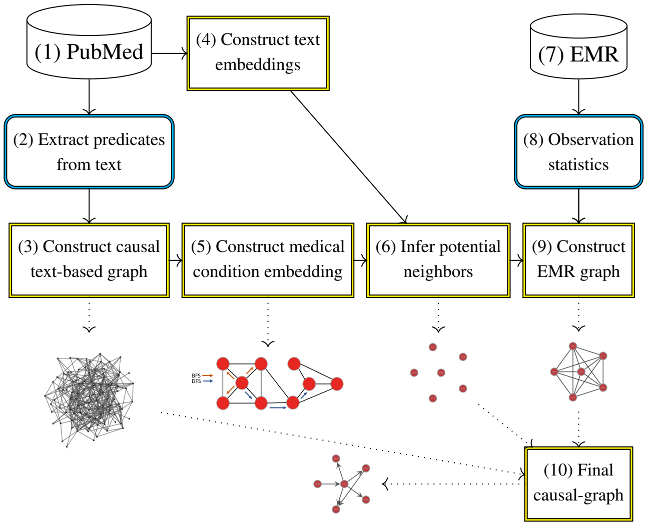

(1) PubMed
(4) Construct text embeddings
(7) EMR
(2) Extract predicates from text
(8) Observation statistics
(3) Construct causal text-based graph
(5) Construct medical condition embedding
(6) Infer potential neighbors
(9) Construct EMR graph
BFS
DFS
(10) Final causal-graph

Figure 1: Construction of the causal-graph

1. Method **C1**: select all causal edges.
2. Method **C2**: select causal edges that occur more than m times; that is, add an edge $e = (d_1, d_2)$ if the PubMed graph contains m or more triples (d_1, p, d_2) where p is causal. Assuming the textual repository might contain some degree of error, this methods sets the support threshold m to filter out singular correlations. In our experiments, we set $m = 2$ for C2.
3. Method **C3**: select causal edges that appear m-times more than non-causal edges between the same terms; that is, if the PudMed graph contains k causal edges (d_1, p, d_2) and k' non-causal edges (d_1, p', d_2), then add an edge $e = (d_1, d_2)$ if $k \geq mk'$. Here, we attempt to reduce noise in the graph by filtering out connections that have stronger non-causal support in the textual repository. In our experiments, we set $m = 2$ for C3.
4. Method **C4**: same as C3, but now with $m = 10$.
5. Method **C5**: select pairs (d_1, d_2) that appear in m edges (d_1, p, d_2), regardless of whether or not p is causal. In our experiments, we set $m = 6$ for C5.

   The parameters were chosen after manual validation on a validation set separate from the test set.

Context-Centric Reduction

We aim to produce a small and precise causal graph, for the sake of both usability by human researchers and automated analysis. As described in Section 1, it is hard to use standard methods for causal inference with large numbers of mediating and confounding variables. We observe that a common use case of medical condition causal graphs is in the context of a specific condition d_0 of interest, where the goal is to infer conclusions relating to d_0. Therefore, we reduce a causal graph G to a subgraph G' of G that is created with respect to a medical condition $d_0 \in \mathbf{D}$ and contains only nodes that are relevant to d_0. To capture *relevance*, we define a proximity function S to measure the relevance of $d_i \in \mathbf{D}$ to d_0, and we restrict the graph to the nodes that are closest to d_0. We explore three alternatives for S: *co-occurrence*, *text embeddings*, and *graph embeddings*.

Co-occurrence. A common assumption in Natural Language Processing (NLP) is that terms that are related appear together often in the text. Following this assumption, we may expect related conditions to appear together in the textual repository. The *co-occurrence* of term d_1 and d_2 is the number of predicates in the PubMed graph in which the two appear together, divided by the total number of predicates they occur in: $\text{co}(d_1, d_2) = \frac{N(d_1,d_2)}{N(d_1)+N(d_2)}$, where $N(d_1, d_2)$ is the number of edges (triples) (d_1, p, d_2) in the PubMed graph, and $N(d_i)$ is the number of triples (d_i, p, d') in the PubMed graph.

Text embeddings. Another common approach is using *word embeddings* based on a textual corpus. Word2vec (Mikolov et al. 2013) is an embedding algorithm which, given an input text corpus, produces a vector representation for each word in the text. It is widely accepted as a baseline for word embeddings in NLP. The distance between word2vec embeddings of two terms can be viewed as the semantic distance between the terms: terms that are closely related are mapped to closer vectors, an vice versa.

Graph embeddings. Considering only semantic proximity does not take into account the relations between terms as they appear in text. For example, consider a corpus composed of the following sentences: “The patient has a family history of asthma and diabetes,” and “Hypertension can lead to diabetes.” Here, the co-occurrence measure of the terms “diabetes” and “asthma” in the first sentence, and the terms “diabetes” and “hypertension” in the second sentence will be the same, although the second sentence portrays a much stronger causal relation. If we take into account the context that these diseases appear in, we can find that diabetes and hypertension are often mentioned as related diseases, while asthma and diabetes are not.

Node2vec (Grover and Leskovec 2016) is an embedding algorithm that generalizes word2vec for the graph domain. The nodes in the graph can be regarded as words and the algorithm creates “sentences” by generating random walks over the graph starting from each node. The algorithm’s hyper-parameters are used to control whether the graph walks they preform are local “within cluster” walks similar to BFS, or more global walks which are more similar to DFS. In the first option, node2vec will produce similar embeddings for nodes in the same cluster. For the second option, nodes that preform a similar structural role in the graph (i.e. connecting node, central node) will have similar embeddings. Given a random walk from node v to node u, node2vec formulates this bias strategy by defining two hyper-parameters, $\tilde{p}$ and $\tilde{q}$, which help adjust the transition probability $\alpha_{\tilde{p}\tilde{q}}(v, x)$ from node v to some node x:

$$\alpha_{\tilde{p}\tilde{q}}(v, x) = \begin{cases} \frac{1}{\tilde{p}} & if \quad d_{vx} = 0 \\ 1 & if \quad d_{vx} = 1 \\ \frac{1}{\tilde{q}} & if \quad d_{vx} = 2 \end{cases}$$

where d_{vx} is the distance between node v and node x.

In this way, node2vec can bias the random walk closer or further away from the source node. This creates different embedding types. Setting $\tilde{p} < \tilde{q}$ biases the random walk to nodes closer to each other. This, in turn, causes nodes from the same cluster to be embedded closer and nodes from different neighborhoods to be embedded further away. Setting $\tilde{p} > \tilde{q}$ biases the random walk to embed nodes of the same graph structural role closer together while others are embedded further away. As node2vec only uses the transition probability, by weighting and directing the random walks, the embedding algorithm can be extended for weighted and directed networks as well. We extend the node2vec algorithm and apply its random walks on a directed and weighted causal text-based graph where the directed edges represent hypothesized causal relations between medical conditions. The edge weights are set according to the co-occurrence of the two conditions in the text. Randomly traversing this graph can be intuitively thought of as the causal paths a *causes* b *causes* c and so on. The embeddings produced by this method capture a disease’s position in the causal structure of the graph, as well as its textual semantic properties (since the graph is based on text), adding context to the resulting causal graph.

Incorporating the EMR Based Graph

Even after using the above text-analysis techniques, the text-based causal graph still contains many weak or uncertain causal connections. In this section, we explain how we leverage EMR data to construct the final causal-graph. While the text graph is based on causal connections implied by natural language in the text, the EMR graph is undirected and contains only correlations. On itself, the EMR graph holds no causal information. We use a *lack of correlation* in the EMR data as a criterion for pruning the dense text-derived graph. We look at all patient diagnosis in the EMR data. In order to establish lack of correlation we preform a pairwise Pearson’s chi-square test with a 95% significance-level testing whether the population of patients diagnosed with d_1 is independent of the populations of patients diagnosed with d_2. If the populations are *independent*, we take it that there is no supporting evidence for the causal relation in the EMR data and we therefore *remove* the corresponding edge $e = (d_1, d_2)$ from the text-based causal graph.

To reduce statistical uncertainty, we only consider medical conditions for which more than 5000 patients were diagnosed (out of over 1.5M patients in the data base), and since we are only looking for correlations, we analyze the data spanning a six year period (2005-2010), looking at all the diagnoses at this time, disregarding their temporal order of appearance.

5 Results

We now describe our experimental study over our techniques for constructing the causal graph.

Setup. We construct causal graphs for two diseases: *Celiac* and *Atopic Dermatitis*. Both are relatively common diseases, are not specific to a certain age or other group, and long-term conditions. From the graphs, we generated a list of

Table 1: Evaluation Example for Atopic Dermatitis

Medical Condition	**Evaluation**
Eczema Herpeticum	Positive
Molluscum Contagiosum	Negative
Allergy to eggs	Negative
Contact Dermatitis	Negative
Dermatitis, Irritant	Negative
Skin disorder	Positive
Dermatitis, Exfoliative	Unknown
Mite Infestations	Unknown
Immediate hypersensitivity	Positive
Allergic Conjunctivitis	Negative
Metal allergy	Negative
Urticaria	Negative

possible causal links by searching for neighbors of distance c from the target disease d_0. The lists were evaluated manually by medical professionals who were given a list of connected medical conditions for each target disease. The evaluators were asked to mark each putatively linked disease on the list as “positive,” “negative” or “do not know.” They were instructed to select “positive” if the condition is either caused by or causing the target disease and “negative” otherwise. If they were unsure they chose “do not know.” An evaluation example is presented in Table 1. We combined the evaluators responses using only medical conditions for which a “positive” classification was agreed upon by all evaluators (Cohen’s kappa > 0.8).

We look at the Positive Predictive Value (PPV), i.e. the ratio between the number of medical conditions classified by the evaluators as “positive” and the total number of medical conditions linked to the target disease in a given graph. We wish to emphasize that for our calculation of PPV we compare our graphs with the knowledge of the medical practitioners, as there is no absolute ground truth in this field and it is very much possible that the physicians’ knowledge is incomplete. Moreover, we claim that new knowledge is likely to be discovered when using large textual and EMR repositories.

As a baseline for comparison, we use *SemCause*: the graph that consists of all edges (d_1, d_2) such that the PubMed graph contains a triple (d_1, p, d_2) where p is a causal predicate (as defined in Section 4).

We conduct our experiment in two stages. First we show that limiting the node set N increases precision of the text-based causal graph. Then we show that creating a merged graph, i.e. combining EMR and textual data, further enhances precision.

PPV Results

For our experiments, we use $c = 5$ to determine a causal connection. That is, if there exists a directed path of length $<= 5$ between $source$ and d_0 or between d_0 and $source$ we add $source$ to the list. We constructed the causal graph using method C1 for constructing textual graph edges.

Table 2: PPV for Celiac by a human evaluator.

Method	Text	Merged
SemCause	0%	0%
word2vec	23%	27%
co-occurrence	24%	35%
node2vec + co-occurance	**32%**	**50%**
node2vec combined	29%	33%
node2vec + word2vec	17%	20%

Due to the size of the SemCause graph, its PPV was evaluated using a randomly sampled set of ten edges. Tables 2 and 3 give the results for Celiac and Atopic Dermatitis, respectively. The tables show the PPV of the PubMed graph compared to the merged graph. The merged graph is the reduced text-based graph incorporated with the EMR based graph as described in Section 4. We see that the PPV for the SemCause sample was zero for both diseases, i.e., none of the connections it suggested were evaluated as valid causal connections by the evaluators.

We first observe that constructing a text-based graph with a smaller group of nodes is empirically better than using a graph induced from SemCause with no node selection. We also note that for Celiac, the merged graphs are smaller but maintain as high or higher PPV. This means that the merging chooses the right edges, which to a degree validates our choice. However for Atopic Dermatitis this is not always the case. Additionally, we observe that the graphs based on the nodes created using node2vec were the most precise for both diseases. This indicates the benefit of using the entire SemCause graph community structure.

An interesting observation is the difference in overall PPV between the two diseases. The graphs for Celiac were overall more precise than the graphs for Atopic Dermatitis, and had much more agreement amongst evaluators. In some cases, the agreement amongst evaluators for the Atopic Dermatitis graph was so low that it led to very low (even zero) PPV results. This might be due to the fact that our evaluators had different knowledge or experience in that specific disease. Atopic Dermatitis is generally less studied than Celiac, therefore the medical practitioners have less knowledge of causal relations for this disease.

Comparing PubMed selection algorithms. As described in Section 4, we create two types of edges: PubMed-based and EMR-based edges. The EMR-based edges are extracted from the EMR data based on a statistical correlation test. We now compare the agreement between the PubMed-based edges according to the different methods, and EMR-based edges.

Table 3: PPV for Atopic Dermatitis by a human evaluator.

Method	Text	Merged
SemCause	0%	0%
word2vec	20%	30%
co-occurrence	7%	0%
node2vec + co-occurance	8%	0%
node2vec combined	29%	25%
node2vec + word2vec	24%	**33%**

Table 4: Comparing precision and recall of the different text edge algorithm compared to a standard of EMR correlation identification. The notation a/b in the table refers to the method the underline graph is built. For example, w2v/co means the underline graph is built using word2vec (w2v) nodes and co-occurrence (co) weights. Results are presented for Celiac disease.

Method	Nodes	Edges	EMR-r	EMR-p
word2vec C1	36	49	0.75	0.65
word2vec C2	36	29	0.05	0.25
word2vec C3	36	39	0.3	0.67
word2vec C4	36	23	0.1	0.5
word2vec C5	36	25	0.6	0.67
co-occurrence C1	37	179	1	0.57
co-occurrence C2	37	113	0.95	0.59
co-occurrence C3	37	166	0.95	0.56
co-occurrence C4	37	112	0.95	0.56
co-occurrence C5	37	84	0.95	0.57
node2vec co/co C1	22	79	1	0.73
node2vec co/co C2	22	56	0.94	0.71
node2vec co/co C3	22	75	0.94	0.71
node2vec co/co C4	22	55	0.81	0.72
node2vec co/co C5	22	40	0.94	0.71
node2vec w2v/co C1	16	29	1	0.8
node2vec w2v/co C2	16	19	0.25	0.75
node2vec w2v/co C3	16	22	0.58	0.78
node2vec w2v/co C4	16	14	0.25	0.75
node2vec w2v/co C5	16	16	0.92	0.85
node2vec w2v/w2v C1	29	33	0.56	0.77
node2vec w2v/w2v C2	29	18	0.17	0.75
node2vec w2v/w2v C3	29	30	0.28	0.71
node2vec w2v/w2v C4	29	3	0.11	0.67
node2vec w2v/w2v C5	29	6	0.22	0.67

We evaluate the different methods for building the text-based graph using two measures: EMR recall and EMR precision. EMR precision is the number of relations that are common to the text-based and EMR-based graphs divided by the number of correlations in the EMR-based graph.

$$\mathrm{EMRp}(\mathrm{G}_{\mathrm{text}}, \mathrm{G}_{\mathrm{EMR}}) = \frac{|\mathrm{Corr}_{\mathrm{EMR}}(\mathrm{target}) \cap \mathrm{Caus}_{\mathrm{text}}(\mathrm{target})|}{|\mathrm{Caus}_{\mathrm{text}}(\mathrm{target})|}$$

EMR recall is the number of relations that are common to the text-based and EMR-based graphs divided by the number of correlations in the EMR-based graph.

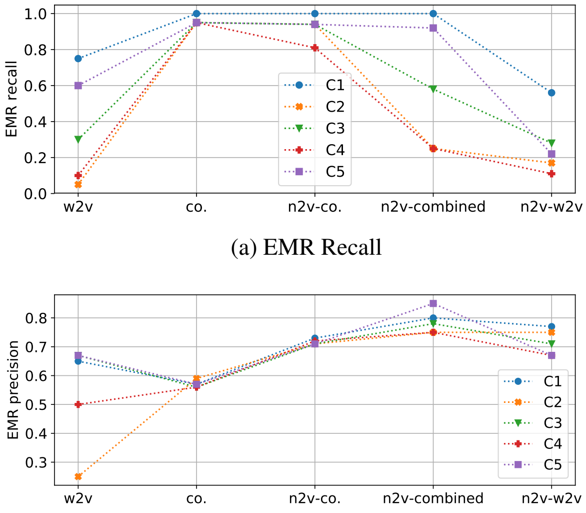
$$\mathrm{EMRr}(\mathrm{G}_{\mathrm{text}}, \mathrm{G}_{\mathrm{EMR}}) = \frac{|\mathrm{Corr}_{\mathrm{EMR}}(\mathrm{target}) \cap \mathrm{Caus}_{\mathrm{text}}(\mathrm{target})|}{|\mathrm{Corr}_{\mathrm{EMR}}(\mathrm{target})|}$$

EMR precision and recall for the Celiac graph built with methods C1 to C5 as described in Section 4 are presented in Table 4 and Figures 2a and 2b. The x-axis represents the different graph composition methods noted in abbreviation: word2vec (w2v), co-occurance (co.), node2vec based on co-occurance (n2v-co.), node2vec based on word2vec and co-occurance combined (n2v combined), and node2vec based on word2vec (n2v-w2v).

As we pose more constraints on edge selections, we built smaller graphs, naturally leading to lower EMR recall numbers. It is also interesting to observe that the co-occurrence method and the node2vec based on co-occurrence methods appear to be less influenced by method chosen, remaining with stable overlap values between EMR and text. we can see in Figure 2a that EMR recall is generally higher for co-occurrence or node2vec based graphs and amongst the node2vec based graphs the EMR recall is higher when node2vec was fitted on a co-occurrence based graph. A similar conclusion can be drawn from Figure 2b regarding the EMR precision. The C1 method seems to be comparable or superior to the other methods and is also the most inclusive. We thus use this method for constructing the text-based graphs.

Contribution of EMR filtering. The text-based graph may contain redundant and erroneous links. The EMR data provides us with important "supporting evidence" that is used for narrowing down the text-based graph.

Figure 3 describes the EMR recall and EMR precision of the text-based graphs created for Celiac and Atopic Dermatitis with methods C1. Table 5 and Table 6 present the number of nodes and edges in each graph. EMR recall and EMR precision give us insight into the amount of overlap or "agreement" between the text-based graph and the EMR data. The more common relations there are, the greater the agreement between the two graphs. We do not necessarily desire complete agreement between the graphs, as the EMR data contains correlations that are not necessarily causal. We seek correlations in the EMR for as many causal relations discovered in the text-based data but not vice versa. Note that the text-based graph created using the node2vec based on word2vec and co-occurrence combined achieves the highest EMR recall and precision. Additionally, it achieves the highest PPV as described in Table 3 and Table 2.



(a) EMR Recall

(b) EMR Precision

Figure 2: Method Performance

Qualitative Example: Celiac

We continue using Celiac as a qualitative example. Celiac disease causes an immune response to gluten consumption

Table 5: Celiac graph data.

Method	Nodes	Edges
SemCause	203	599
word2vec	36	49
co-occurrence	37	179
node2vec + co-occurrence	22	79
node2vec combined	16	29
node2vec + word2vec	29	33

Table 6: Atopic Dermatitis Graph Data

Method	Nodes	Edges
SemCause	143	396
word2vec	34	68
co-occurrence	32	132
node2vec + co-occurrence	14	27
node2vec combined	16	42
node2Vec + word2vec	31	63

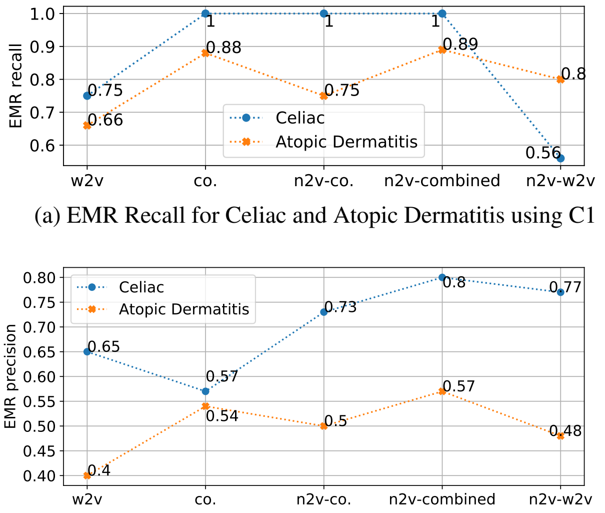
which damages the lining of the small intestines (Booth 1977). An immediate result of this damage is malabsurbtion of nutrients. The medical evaluators annotated medical conditions related to malabsorption and malnutrition as correct causal relations. Some digestive system inflammatory conditions were also identified. Other terms, such as other autoimmune diseases, Pancriatic insufficiency, Downs Syndrome and Dermatitis Herpetiformis were noted by them to be correlative with Celiac, but without a known causal relation. The merged graph created for the celiac disease is shown in Figure 4.

Dermatitis Herpetiformis is a skin condition that is expressed as a rash in response to gluten ingestion. Some explanations of the causal relation between the two conditions also exist.[2] Our text-based causal graph contains direct causal edges between the terms. Exploring the graph we see the following causal paths:

Celiac → Steatorrhea → Malnutrition → Diabetes Mellitus
Celiac → Malabsorption Syndromes → Malnutrition
→ Diabetes Mellitus

The first part of the paths, stating that Celiac causes malnutrition or malabsorption, was acknowledged as a viable causal connection. The second part, stating that malnutrition may cause diabetes, is not an irrational induction. Inspecting the causal paths between Downs Syndrome and Celiac disease. We find several paths that originate from the predicate “Predisposes” which has a weaker causal meaning and is perhaps incorrect in this case.

Although our methodology will not create new causal connections if none existed in the original medical literature repository, based on the feedback we received from our clinical evaluators, we believe it may support novel discoveries as it is helpful in bringing forward information that is known but has not been considered as an explanation of a particular phenomena.



(a) EMR Recall for Celiac and Atopic Dermatitis using C1

(b) EMR Precision for Celiac and Atopic Dermatitis using C1

Figure 3: EMR precision and EMR recall for test cases

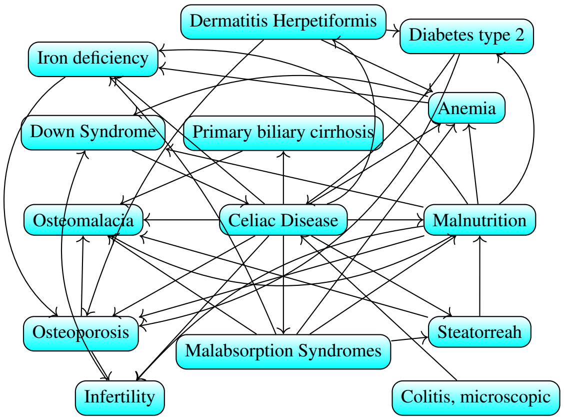


Figure 4: Merged causal graph for celiac disease

6 Conclusions

In this paper, we suggest several methods for constructing causal medical condition graphs. They are of value in many applications, such as clinical trial design, where potential confounders need to be identified and adjusted for. We apply state-of-the-art natural-language-processing techniques for extracting a small and relevant set of diseases for the causal graph, and further prune this graph using correlations found in EMR data. In this manner we incorporate the EMR data, which does not contain causal information, with causal relations extracted from medical literature. The resulting graph is both useful and relatively precise, as assessed by experts.

Although the task is of high importance for medical research, there is currently no benchmark which can be used as evaluation of such work. In our work, we compared the resulting graphs to domain experts. We consider our work as the first step for a causality database, which is open for other researchers for use and expansion and can serve as a

[2] https://www.niddk.nih.gov/health-information/digestive-diseases/dermatitis-herpetiformis

benchmark. We believe an automated method for extracting knowledge based on large theoretical and observational resources, such as the one we presented, will bring high value to the medical community and enable further research.

7 Acknowledgments

This work was supported in part by the Lorry I. Lokey Interdisciplinary Center for Life Sciences and Engineering.

References

Avillach, P.; Dufour, J.-C.; Diallo, G.; Salvo, F.; Joubert, M.; Thiessard, F.; Mougin, F.; Trifirò, G.; Fourrier-Réglat, A.; Pariente, A.; et al. 2012. Design and validation of an automated method to detect known adverse drug reactions in medline: a contribution from the eu–adr project. *Journal of the American Medical Informatics Association* 20(3):446–452.

Bodenreider, O. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research* 32(suppl1):D267–D270.

Booth, C. C. 1977. Coeliac disease. *Nutr Metab* 21(1-3):65–75.

Bui, Q.-C.; Nualláin, B. Ó.; Boucher, C. A.; and Sloot, P. M. 2010. Extracting causal relations on hiv drug resistance from literature. *BMC Bioinformatics* 11(1):101.

Casucci, S.; Lin, L.; Hewner, S.; and Nikolaev, A. 2017. Estimating the causal effects of chronic disease combinations on 30-day hospital readmissions based on observational medicaid data. *Journal of the American Medical Informatics Association*.

Claassen, T., and Heskes, T. 2012. A bayesian approach to constraint based causal inference. In *UAI*, 207–216. AUAI Press.

D'Amour, A.; Ding, P.; Feller, A.; Lei, L.; and Sekhon, J. 2017. Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*.

Finlayson, S. G.; LePendu, P.; and Shah, N. H. 2014. Building the graph of medicine from millions of clinical narratives. In *Scientific data*.

Goodwin, T., and M. Harabagiu, S. 2013. Automatic generation of a qualified medical knowledge graph and its usage for retrieving patient cohorts from electronic medical records.

Gottlieb, A.; Yanover, C.; Cahan, A.; and Goldschmidt, Y. 2017. Estimating the effects of second-line therapy for type 2 diabetes mellitus: retrospective cohort study. *BMJ Open Diabetes Research and Care* 5(1).

Grover, A., and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Hauser, A., and Bühlmann, P. 2015. Jointly interventional and observational data: estimation of interventional markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77(1):291–318.

Kilicoglu, H.; Shin, D.; Fiszman, M.; Rosemblat, G.; and Rindflesch, T. C. 2012. Semmeddb: a pubmed-scale repository of biomedical semantic predications. *Bioinformatics* 28(23):3158–3160.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, 3111–3119.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 701–710. ACM.

Radinsky, K., and Davidovich, S. 2012. Learning to predict from textual data. *J. Artif. Int. Res.* 45(1):641–684.

Rindflesch, T. C., and Fiszman, M. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics* 36(6):462 – 477.

Rotmensch, M.; Halpern, Y.; Tlimat, A.; Horng, S.; and Sontag, D. 2017. Learning a health knowledge graph from electronic medical records. *Scientific Reports* 7(1):5994.

Sachs, K.; Perez, O.; Pe'er, D.; Lauffenburger, D. A.; and Nolan, G. P. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. 308(5721):523–529.

Stuart, E. A.; DuGoff, E.; Abrams, M.; Salkever, D.; and Steinwachs, D. 2013. Estimating causal effects in observational studies using electronic health data: challenges and (some) solutions. *Egems* 1(3).

Tang, J.; Qu, M.; Wang, M.; Zhang, M.; Yan, J.; and Mei, Q. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, 1067–1077. International World Wide Web Conferences Steering Committee.

Triantafillou, S., and Tsamardinos, I. 2015. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research* 16:2147–2205.

Triantafillou, S.; Lagani, V.; Heinze-Deml, C.; Schmidt, A.; Tegner, J.; and Tsamardinos, I. 2017. Predicting causal relationships from biological data: Applying automated causal discovery on mass cytometry data of human immune cells. *Sci Rep* 7:12724.